

# Apport des approches phylogénétiques pour expliquer l'origine des génomes mosaïques, exemple chez le Riz

Charles-Elie Rabier <sup>1</sup>, Vincent Berry, Fabio Pardi and Céline Scornavacca

<sup>1</sup>LIRMM, Montpellier, France

charles-elie.rabieratumontpellier.fr

Les réseaux phylogénétiques retranscrivent l'histoire évolutive de groupes d'individus (espèces ou populations), comportant des événements de réticulation comme les phénomènes d'hybridation ou les transferts horizontaux de gènes (e.g. Solis Lemus, Ane 2016). Dans ce contexte, au sein du projet GenomeHarvest (collaboration avec le CIRAD), nous proposons une nouvelle méthode Bayésienne d'inférence de réseau phylogénétique directement depuis les alignements de séquences d'ADN. Les applications portent sur le génome du riz pour lequel plusieurs thèses s'opposent quant au processus de domestication

D'une manière générale, notre méthode se veut une généralisation de SNAPP (Bryant et al, 2012): elle repose sur une modélisation mathématique assez fine et sur des développements algorithmiques permettant le traitement d'une grande masse de données. Les modèles mathématiques étudiés sont des modèles hiérarchiques avec plusieurs niveaux stochastiques. Un processus de coalescence modélise l'évolution d'une famille de gènes à l'intérieur d'un réseau. En particulier, le modèle "multispecies coalescent" modélise le tri de lignées incomplet (ILS), une des sources principales de conflits entre arbres de gènes. A chaque noeud de réticulation du réseau (i.e. noeud possédant 2 parents), une lignée génétique hérite du matériel génétique d'un de ces parents, avec une certaine probabilité (Yu et al. 2012). Le deuxième niveau stochastique s'avère le suivant. Conditionnellement à l'arbre généalogique de la famille de gènes (paramètre non observé), un processus de Markov est utilisé pour modéliser les mutations dans la séquence d'ADN le long de chaque branche de l'arbre.

Comme souligné dans Hohna et al. (2014, 2016), il existe un lien entre modèles phylogénétiques et modèles graphiques. Une idée principale repose sur le découpage de gros modèles probabilistes en composantes représentant des distributions de probabilités conditionnellement indépendantes. A titre d'exemple, on peut citer le logiciel CAFE (De Bie et al., 2006), qui permet le calcul de la vraisemblance de la taille d'une famille de gène à l'aide d'un algorithme, où les nœuds de l'arbre phylogénétique sont traités par « post-order ».

Dans notre étude, dans l'optique du calcul de vraisemblance, nous cherchons à parcourir le réseau de la façon la plus efficace possible, en tenant compte de la non indépendance de certaines branches, en raison des phénomènes de réticulation. Par conséquent, des algorithmes dédiés aux réseaux sont proposés. Notre méthode est implémentée en JAVA dans le logiciel BEAST, très employé dans la communauté évolution.